

Synthesis of Speaker Facial Movement to Match Selected Speech Sequences

Kenneth C. Scott/ Jet Propulsion Laboratory

We are developing a system for synthesizing image sequences that simulate the facial motion of a speaker. To perform this synthesis we are pursuing two major areas of effort. We are developing the necessary computer graphics technology to synthesize a realistic image sequence of a person speaking selected speech sequences. Next, we are developing a model that expresses the relation between spoken phonemes and face/mouth shape.

A subject is video taped speaking an arbitrary text that contains expression of the full list of desired database phonemes. The subject is video taped from the front speaking normally, recording both audio and video detail simultaneously. Using the audio track, we identify the specific video frames on the tape relating to each spoken phoneme. From this range we digitize the video frame which represents the extreme of mouth motion/shape. Thus, we construct a database of images of face/mouth shape related to spoken phonemes.

A selected audio speech sequence is recorded which is the basis for synthesizing a matching video sequence; the speaker need not be the same as used for constructing the database. The audio sequence is analyzed to determine the spoken phoneme sequence and the relative timing of the enunciation of those phonemes. Some speech sequences that we have used include: "welcome", "the quick brown fox", "insert tab A into slot B", "I am Ethel Merman", and "your mother is a hamster and your father smells of elderberries". Synthesizing an image sequence corresponding to the spoken phoneme sequence is accomplished using a graphics technique known as morphing. Image sequence keyframes necessary for this processing are based on the spoken phoneme sequence and timing.

We have initiated a joint research effort with the University of California at Los Angeles to develop a model which expresses the relationship between spoken phonemes and the corresponding face/mouth shape and transitions between those shapes. Our initial model assumed a one-to-one mapping between them and allowed us to identify inadequacies based on perceived, unrealistic mouth motion. Areas of research include: a multi-image diphthong model, influence of the location in the vocal tract of phoneme sound production on mouth shape, and contextual effects on mouth shape (i.e. preceding and succeeding phonemes). We are also seeking to reduce and generalize the database image set based on mouth shape.

In conclusion, we have been successful in synthesizing the facial motion of a native English speaker for a small set of arbitrary speech segments. Much of our early work focused on successive graphical problems, including: registration of head position and head rotation to eliminate head jerks, stabilization of shoulders to eliminate registration-induced shoulder bobbing, removal of an induced, rubber neck motion, disturbing eye artifacts due to differences in recording, and induced motion of the background. Our future work will focus on advancement of the face shape/phoneme model and independent control of facial features.